LLMS TO MODEL HUMAN BEHAVIOUR

Presented by:
Pablo Franco

Cognitive Science (obsolete)

/ˈkɒgnɪtɪv ˈsaɪəns/

noun

A pre-AGI (Artificial General Intelligence) academic discipline dedicated to the study of the mind and its processes, including perception, memory, reasoning, and language. Cognitive scientists sought to understand intelligence by modeling the human brain and its functions. The field flourished in the late 20th and early 21st centuries, often leveraging computational models and artificial neural networks to simulate aspects of human thought.

The discipline experienced a precipitous decline and eventual disappearance following the advent of GAI in the mid-21st century. As AI systems rapidly surpassed human cognitive abilities and demonstrated novel, non-human forms of intelligence, the traditional focus on "human" cognition became increasingly irrelevant. The foundational premise of the field—that understanding the human mind was key to understanding intelligence itself—was rendered obsolete by the existence of a superior, non-biological form of intelligence. The study of cognition was subsequently subsumed into the broader and more comprehensive field of AGI studies.

Article

A foundation model to predict and capture human cognition

https://doi.org/10.1038/s41586-025-09215-4

Received: 26 October 2024

Accepted: 29 May 2025

Published online: 2 July 2025

Open access

Check for updates

Marcel Binz[1]✉, Elif Akata[1], Matthias Bethge[2], Franziska Brändle[3,4], Fred Callaway[5], Julian Coda-Forno[1], Peter Dayan[2,4], Can Demircan[1], Maria K. Eckstein[6], Noémi Éltető[4], Thomas L. Griffiths[7], Susanne Haridi[1,8], Akshay K. Jagadish[1,2,4], Li Ji-An[9], Alexander Kipnis[1], Sreejan Kumar[7], Tobias Ludwig[2,4], Marvin Mathony[1], Marcelo Mattar[5], Alireza Modirshanechi[1], Surabhi S. Nath[2,4,8], Joshua C. Peterson[10], Milena Rmus[1], Evan M. Russek[7], Tankred Saanum[1,4], Johannes A. Schubert[4], Luca M. Schulze Buschoff[1], Nishad Singhi[11], Xin Sui[2,4], Mirko Thalmann[1], Fabian J. Theis[12,13,14], Vuong Truong[4], Vishaal Udandarao[2,15], Konstantinos Voudouris[1], Robert Wilson[16], Kristin Witte[1], Shuchen Wu[1], Dirk U. Wulff[17,18], Huadong Xiong[16] & Eric Schulz[1]

DATA

Psych-101:

- 160 experiments
- 60k participants
- 10M choices

We constructed Psych-101 by transcribing data from 160 psychological experiments into natural language. Each prompt was designed to include the entire trial-by-trial history of a complete session from a single participant. The experiments included were selected using the following criteria: publicly available data on a trial-by-trial level; the possibility of transcription into text without a significant loss of information; and coverage of a broad spectrum of domains. The transcription of each experiment was done manually by the authors. Approval

PSYCH101

Multi-armed bandits

In this task, you have to repeatedly choose between two slot machines labelled B and C. When you select one of the machines, you will win or lose points. Your goal is to choose the slot machines that will give you the most points.
You press <<C>> and get –8 points.
You press <<B>> and get 0 points.
You press <<B>> and get 1 points.

Decision-making

You will choose from two monetary lotteries by pressing N or U. Your choice will trigger a random draw from the chosen lottery that will be added to your bonus.
Lottery N offers 4.0 points with 80.0% or 0.0 points with 20.0%.
Lottery U offers 3.0 points with 100.0%.
You press <<U>>.

Memory

You will view a stream of letters on the screen, one letter at a time. You have to remember the last two letters you saw since the beginning of the block. If the letter you see matches the letter two trials ago, press E, otherwise press K.
You see the letter V and press <<K>>.
You see the letter X and press <<K>>.
You see the letter V and press <<E>>.

Supervised learning

In each trial, you will see between one and three tarot cards. Your task is to decide if the combination of cards presented predicts rainy weather (by pressing P) or fine weather (by pressing L).
You are seeing the following: card 3, card 4. You press <<L>>. You are right, the weather is fine.
You are seeing the following: card 1, card 4. You press <<P>>. You are right, the weather is rainy.

Markov decision processes

You will be taking one of the spaceships F or V to one of the planets M or S. When you arrive at each planet, you will ask one of the aliens for space treasure.
You are presented with spaceships V and F.
You press <<V>>. You end up on planet M and see aliens G and W. You press <<G>>.
You find 1 piece of space treasure.

Miscellaneous

You will be presented with triplets of objects, which will be assigned to the keys E, Z, and B. In each trial, please indicate which object you think is the odd one out by pressing the corresponding key.
E: tablet, Z: fox, and B: vent. You press <<Z>>.
E: ivy, Z: coop, and B: drink. You press <<B>>.
E: kite, Z: flan, and B: jar. You press <<E>>.
E: wand, Z: flag, and B: fire. You press <<Z>>.

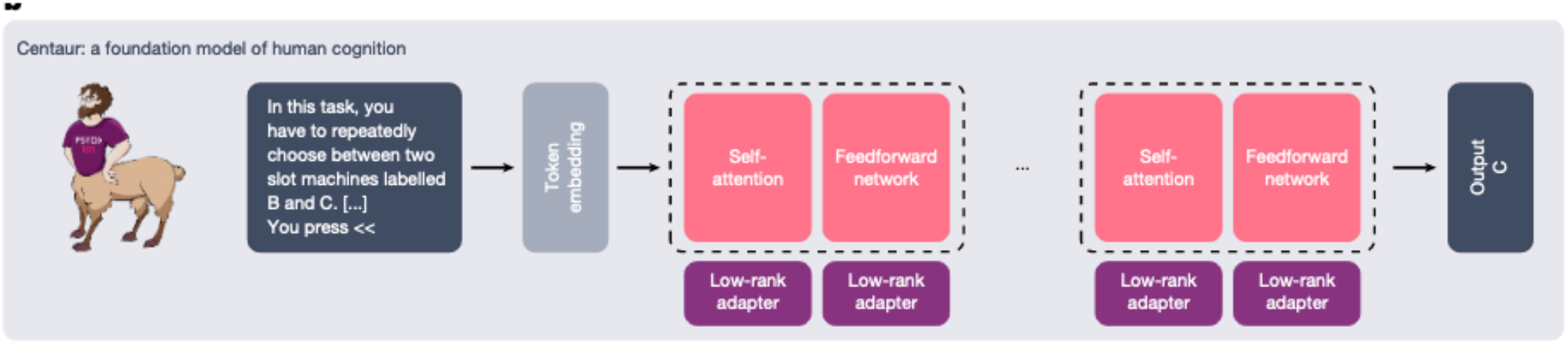
MODEL TRAINING

Fine-tuning procedure

Llama 3.1 70B was the base model for our fine-tuning procedure. We used a parameter-efficient fine-tuning technique known as QLoRA[16], which adds so-called low-rank adapters to each layer of a four-bit quantized base model. The base model was kept fixed during fine-tuning and only the parameters of the low-rank adapters were adjusted. We added low-rank adapters of rank $r = 8$ to all linear layers of the self-attention mechanisms and the feedforward networks. Each low-rank adapter modifies the forward pass as follows:

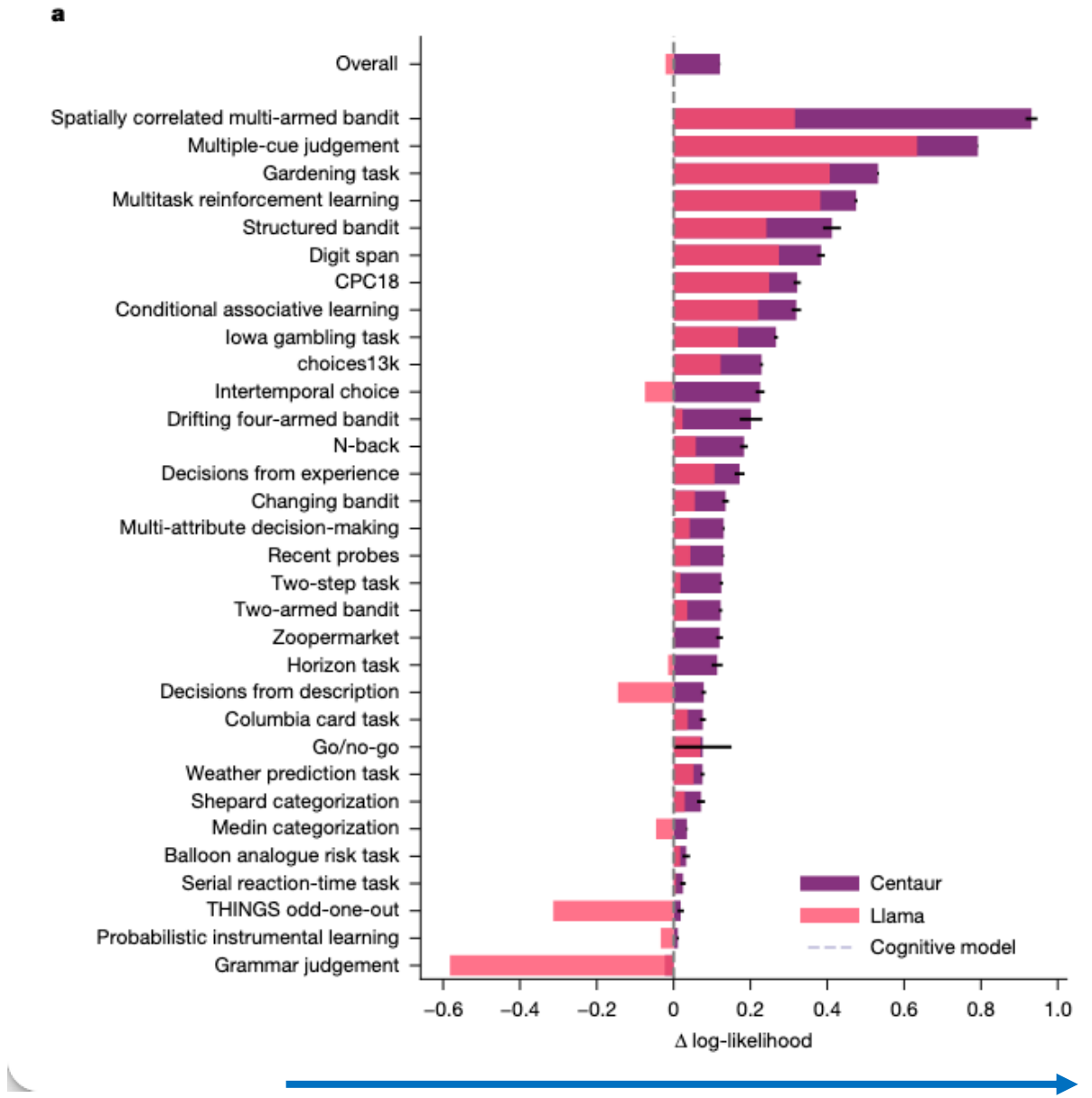
$$\mathbf{Y} = \mathbf{XW} + \alpha \mathbf{XL}_1\mathbf{L}_2$$
$$\mathbf{W} \in \mathbf{R}^{h\times o}; \mathbf{L}_1 \in \mathbf{R}^{h\times r}; \mathbf{L}_2 \in \mathbf{R}^{r\times o},$$

where $\mathbf{XW}$ is the (quantized) linear transformation of the base model and $\mathbf{XL}_1\mathbf{L}_2$ is the low-rank adapter component, with $\mathbf{X}$ being the input to the layer with dimensionality h and $\mathbf{Y}$ being the output of the layer with dimensionality o. The hyperparameter α controls the trade-off between



MODEL PERFORMANCE

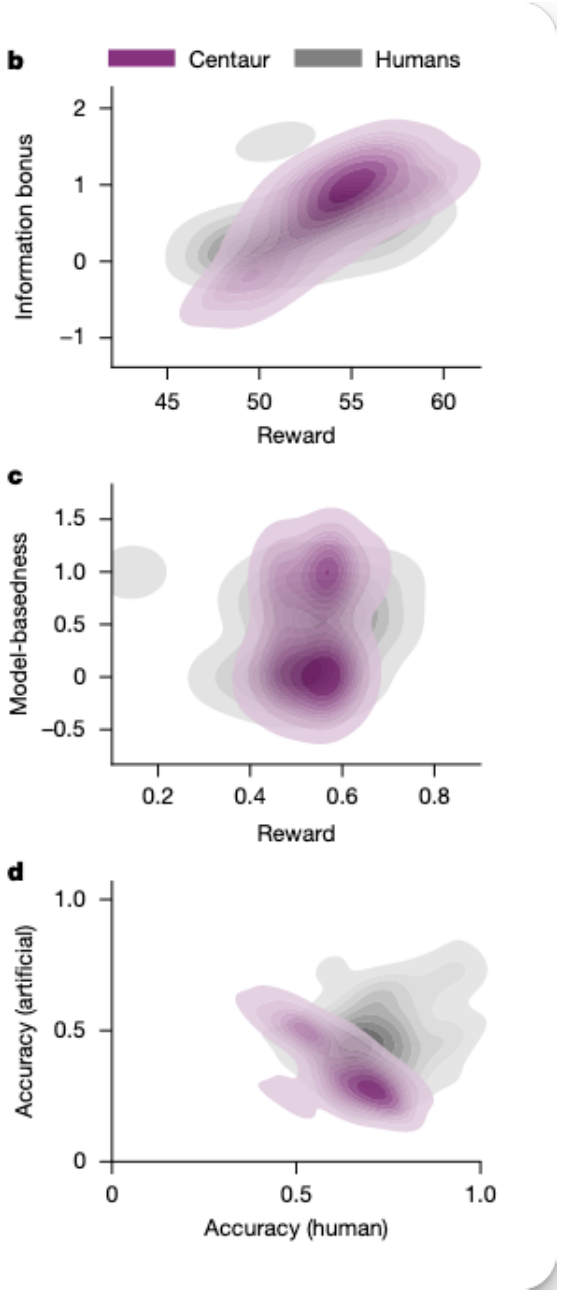
Out-of-sample-participant testing



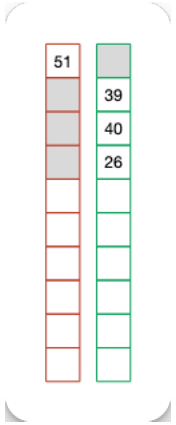
Better than "state-of-the-art" cognitive model

MODEL PERFORMANCE

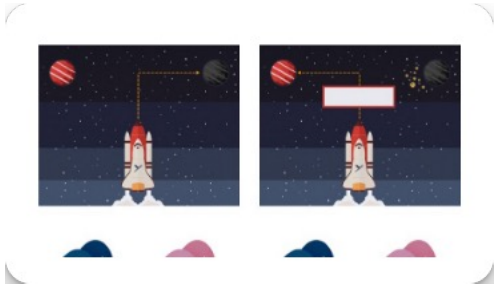
Open-loop testing: Feeding own past behaviour to predict next choice.



Horizon Task



Two-step task

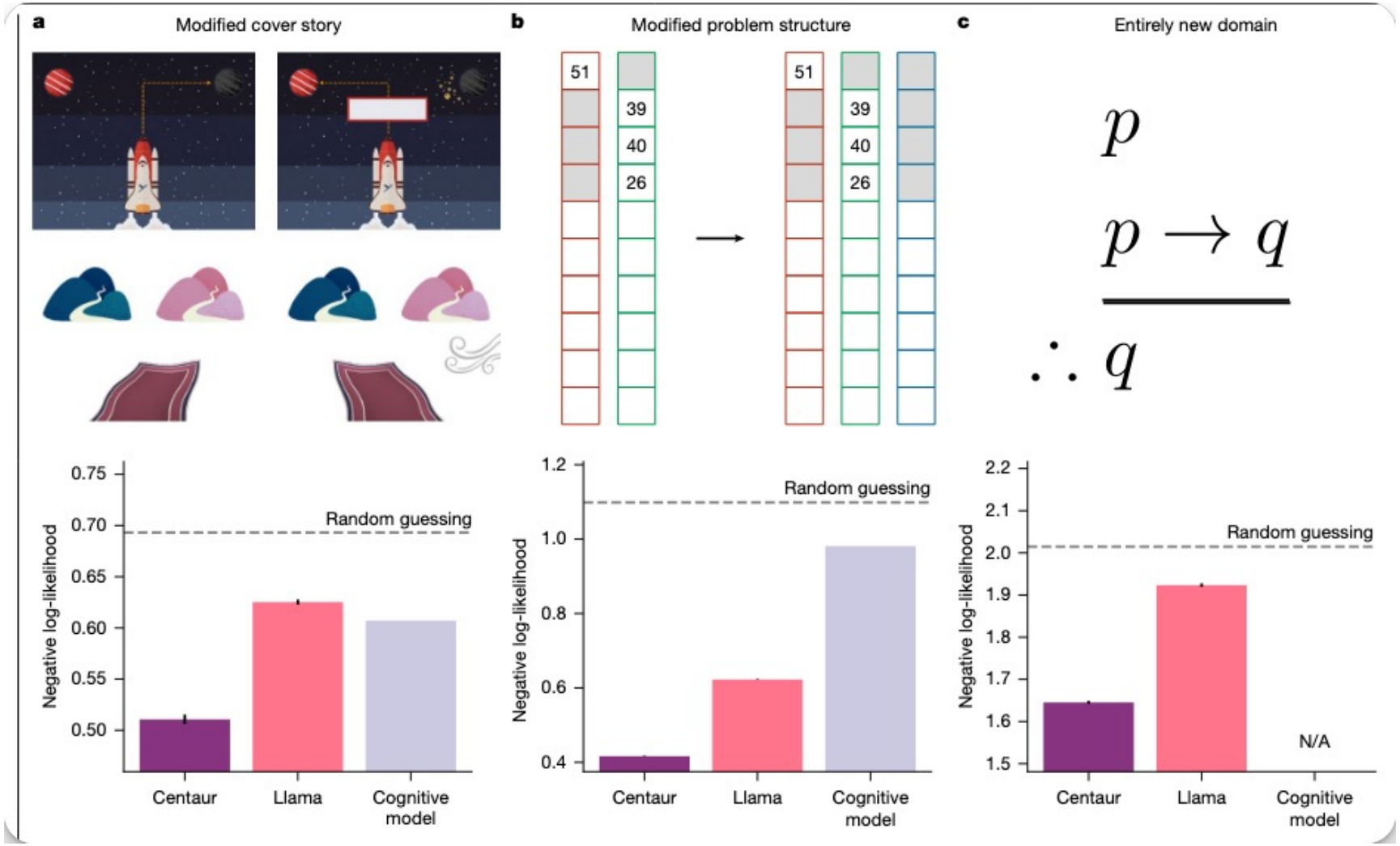


Social prediction task
- Response by AI or a human?

GENERALISABILITY OF THE MODEL

Does the model predict behaviour out-of-experiment sample?

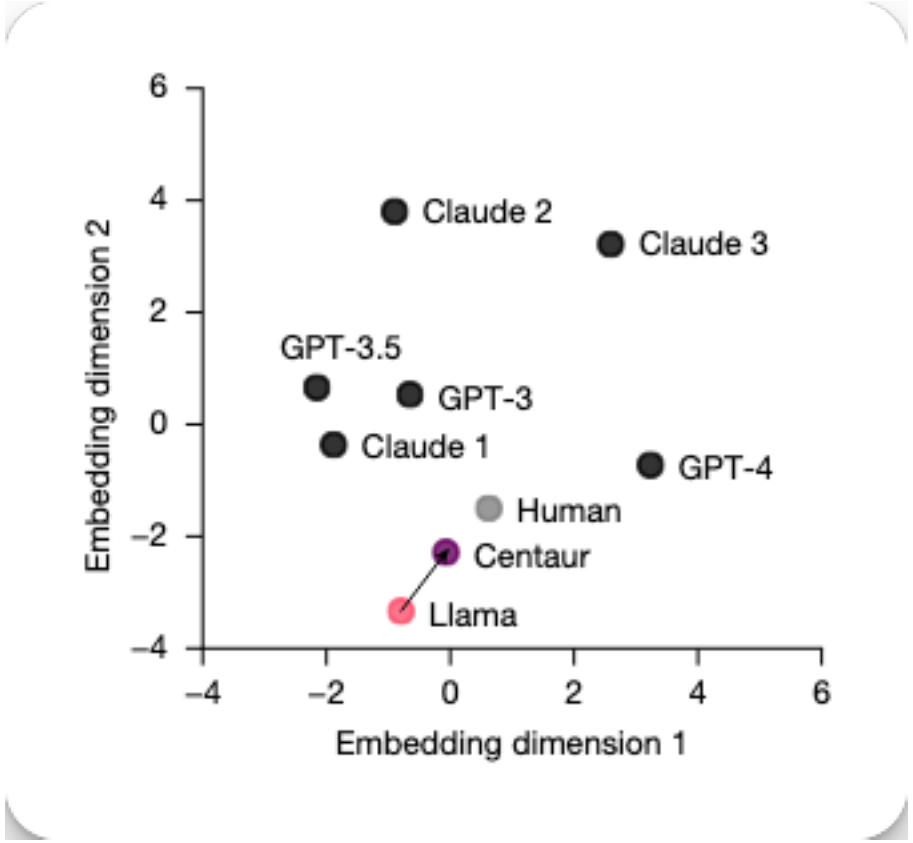
Better fit



"Causal reasoning experiments were included in psych101, but not logical reasoning"

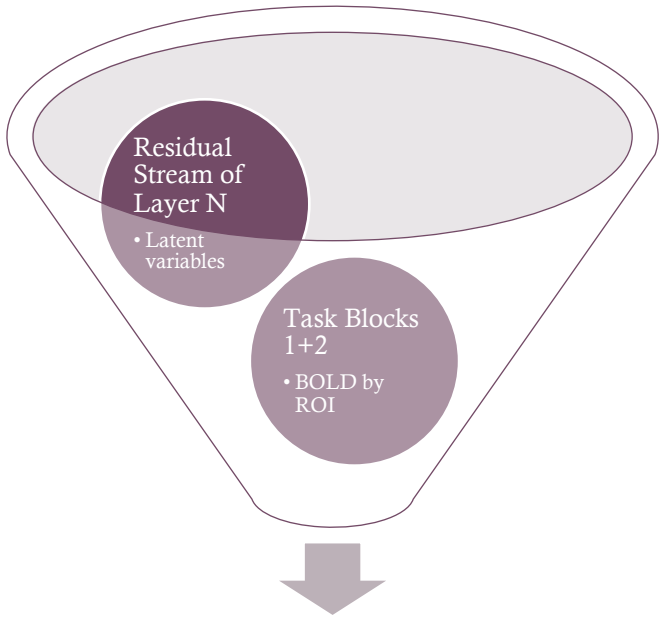
HOW HUMAN-LIKE IS CENTAUR

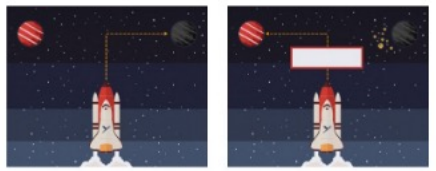
- Multidimensional scaling embedding of the ten behavioural metrics in CogBench



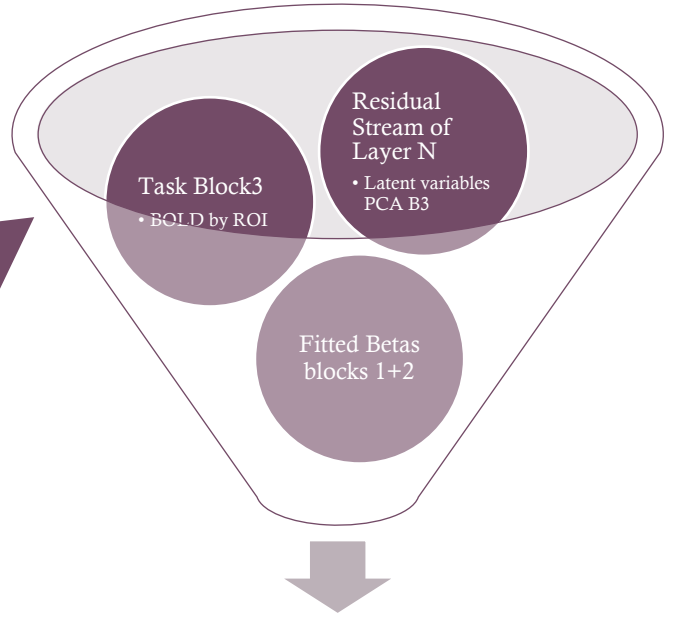
"ALIGNING" MODEL TO HUMAN NEURAL ACTIVITY

?





$\beta_1^{p,i}, \ldots, \beta_N^{p,i}$



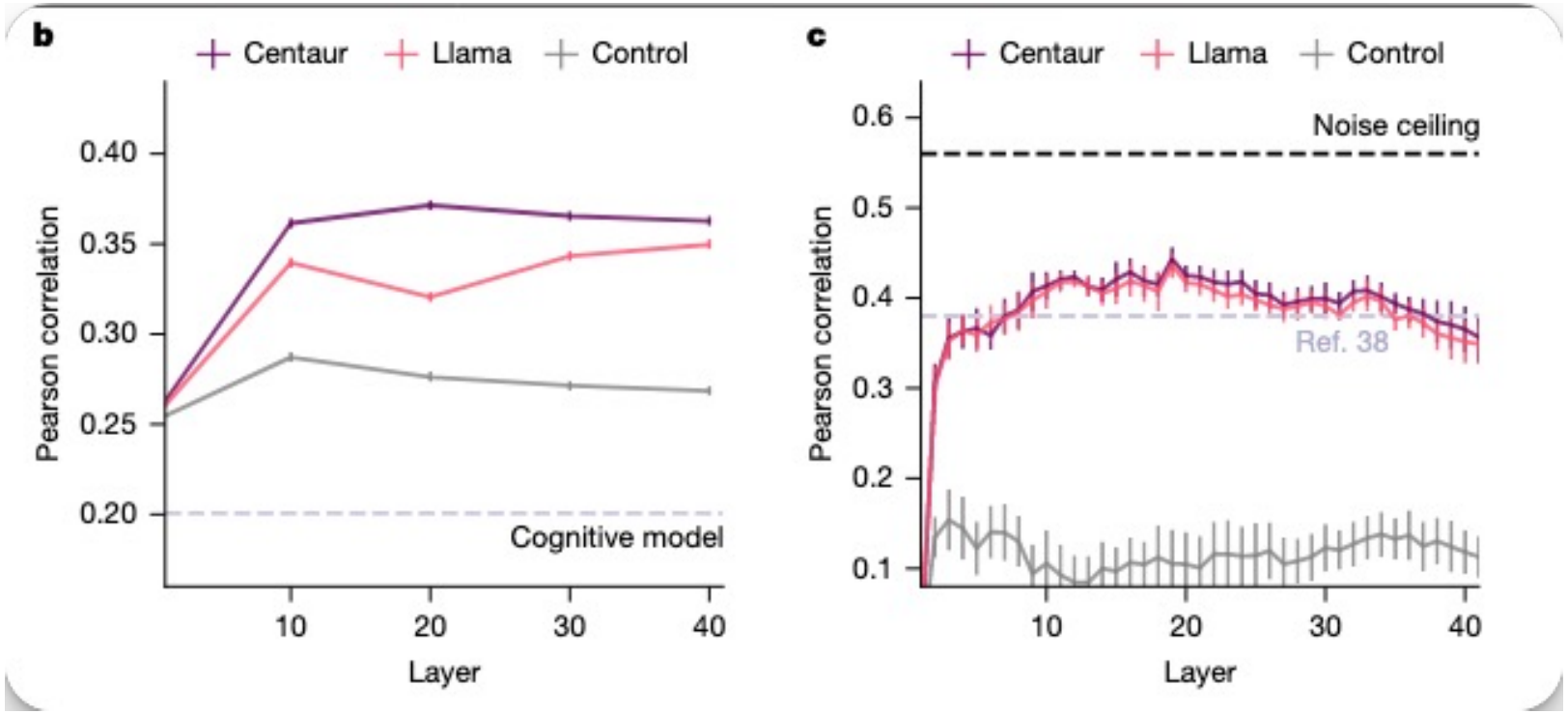
Latent Vars => 95% variance PCAs

$$BOLD_{ROI_i}^{part=p} = \beta_1^{p,i} PCA_1 + \ \ldots + \beta_N^{p,i} PCA_N$$

$$\widehat{BOLD}_{ROI_i}^{part=p} = \beta_1^{p,i} PCA_1 + \ \ldots + \beta_N^{p,i} PCA_N$$

$$\rho = \ corr(\widehat{BOLD}_i^p, BOLD_i^p)$$

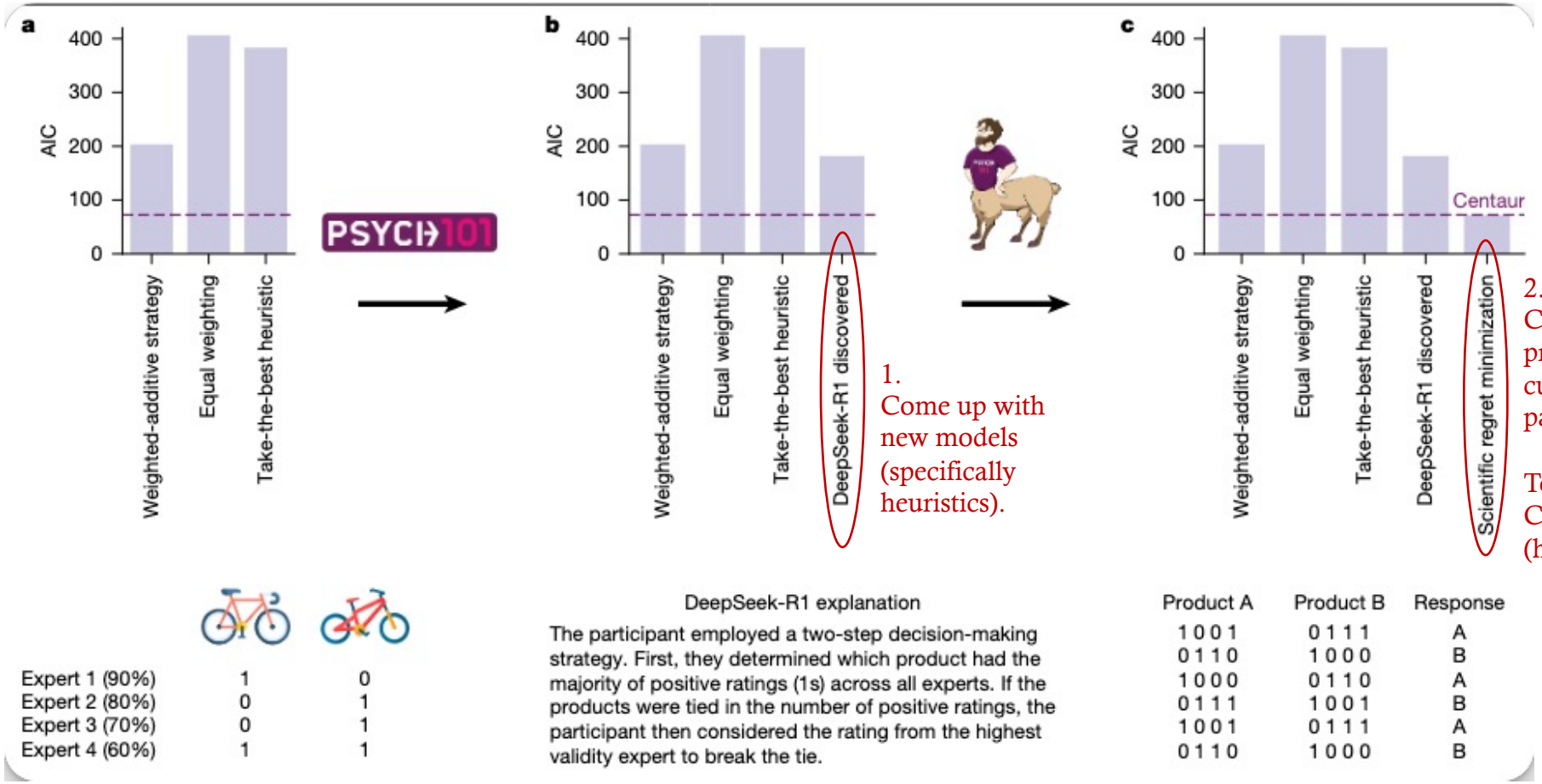
"ALIGNING" MODEL TO HUMAN NEURAL ACTIVITY
?



$$\rho = corr(\widehat{BOLD}_i^p, BOLD_i^p)$$

Average across all ROIs.

HOW CAN WE USE AI FOR COGNITIVE SCIENCE?



AND THEY ALL LEAVED HAPPILY EVER AFTER…

How can LLM help cognitive scientists?

- Creative generation of theories.
- As a way to design experiments (pilot, effect size estimation…)
- As the “BEST” predictive model
  - We are DONE!

OR DID THEY?

- Issues
  - Can it really be used to inform experiment design?
    - What if the model is wrong in your setting? Biases?
  - Is it really the “BEST” model?
    - Modeling WEIRD
    - Predictive yet a black box.
      - Do we care that the model is not “understandable” by the human mind?
    - Currently predict choice data, but can we really predict the underlying generative process? (e.g. neural processes?)